心理科学进展 2023, Vol. 31, No. 10, 1966–1980
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2023.01966

·研究方法(Research Method)·

测验模式效应：来源、检测与应用*

陈 平 代 艺 黄颖诗

(北京师范大学中国基础教育质量监测协同创新中心，北京 100875)

摘 要 测验模式效应(Test Mode Effect, TME)是指同一测验采用不同测验形式施测而产生的测验功能差异。TME 的存在会对测验公平、选拔标准和测验等值等产生影响，因此对 TME 进行准确检测和合理解释具有重要意义。通过对 TME 的来源、检测(包括实验设计和检测方法)以及研究结果进行系统梳理，全面展示 TME 研究的方法论。对 TME 模型进行进一步解释、对 TME 研究中的测验形式进行拓展以及将 TME 的研究成果应用于我国的大规模教育测评项目，都是 TME 领域的未来重要发展方向。

关键词 测验模式效应，测验公平，测量不变性，计算机测验

分类号 B841

1 引言

随着计算机技术的进步和网络的普及，计算机测验已经在测量和评估领域得到广泛使用。大到国际大规模测评项目，小到课堂测试，都越来越多地使用计算机进行施测。测验形式正经历着从传统“纸笔测验(Paper-based Testing, PBT)”向“计算机测验(Computer-based Testing, CBT)”的转变。与 PBT 相比，CBT 具有很多优点，比如：(1)采用计算机辅助测验，测量更加高效、公平；(2)可以呈现高生态效度和高交互性的新颖题型，增加被试的作答兴趣(Pomplun et al., 2006)；(3)可以方便记录被试的作答步骤、动作序列和作答时间等过程性信息，从而更全面地评价被试。正因如此，CBT 已在国际学生评估项目(Programme for International Student Assessment, PISA)、国际数学和科学趋势研究(Trends in International Mathematics and Science Study, TIMSS)、美国国家教育进展评估(National Assessment of Educational Progress, NAEP)等大规模测评项目中得到广泛应用(檀慧玲 等, 2018)。

尽管大多数测验都在朝着 CBT 的方向发展，但这并非一个简单的过程。在进行测验形式的转化之前，研究者和实践者面临一个关键性问题：当同一测验采用不同测验形式(比如 PBT 和 CBT)施测时，其测验结果不一定相同，因而不能盲目地对它们进行直接比较(Jerrim, 2016)。这种由测验形式不同而带来的测验功能差异，被称为测验模式效应(Test Mode Effect, TME; Kroehne et al., 2019; OECD, 2017)。在已有研究中，TME 在绝大多数情况下特指 PBT 和 CBT 这两种测验形式间的差异。考虑到测验形式从 PBT 向 CBT 转变是大势所趋，因此对 TME 进行研究具有以下重要意义：

首先，对 TME 进行研究可以促进测验公平。测验公平是衡量测验质量的一个重要方面，一直受到测验开发者、使用者、心理测量学家和普通大众的广泛关注(Kline, 2013)。一个公平的测验应该能给被试提供平等的机会，来反映他们掌握的与测验目的相关的知识和技能。然而，不同测验形式间的转换可能会引入与测验目的无关的变量，比如被试操作计算机的能力可能会对其CBT的成绩产生影响。因此，研究 TME 有助于明确和控制无关因素的影响，从而提高测验的公平性。

其次，对 TME 进行研究可以保障选拔结果的可比性。很多大型考试都曾出现同时使用 PBT 和

收稿日期：2023-01-10
* 国家自然科学基金面上项目(32071092)、北京师范大学中国基础教育质量监测协同创新中心自主课题(2022-01-082-BZK01)资助。
通信作者：陈平, E-mail: pchen@bnu.edu.cn

CBT 的情况。比如, TOEFL 就同时存在 PBT 和基于互联网的测验(Internet-based Testing, iBT)等多种测验形式。考虑到 TME 的存在, 美国教育考试服务中心在使用 PBT 时, 并不是将 CBT 中的题目直接转移到 PBT 上, 而是有针对性地对 PBT 中的测验内容、实施过程和评价标准等进行修改, 以保障不同测验形式下的结果具有可比性, 从而增加选拔与评价结果的可信度。

最后, 对 TME 进行研究可以帮助获得准确的等值结果。随着 CBT 的广泛使用, PISA 等国际测评项目已经出现“不同测验周期使用不同测验形式”的情况(Feskens et al., 2019)。TME 的存在会影响不同测验周期学生分数等值结果的准确性, 使得研究者没法合理刻画学生的能力发展趋势, 进而削弱教育评估项目的意义。因此, 对题库中可能存在 TME 的题目进行检测, 可进一步改善测验和题目质量, 从而保障教育评估项目的有效性。

鉴于这一主题的重要性, 本文对 TME 进行系统述评, 以期为测量研究者与实践者了解 TME 的来源、检测方法和研究思路提供帮助。本文将按以下顺序进行组织: 首先介绍 TME 的来源, 然后探讨 TME 的检测(包括控制 TME 影响的实验设计和对 TME 进行检测的方法), 接着总结 TME 研究的结果与不足, 最后展望 TME 的未来研究方向。

2　TME 的来源

TME 来源于测验形式不同所带来的差异, 这种差异可以来自 4 个层面: 测验层面、题目层面、被试层面和评分者层面。接下来分别介绍这 4 个层面的差异如何导致 TME 的产生。

2.1　测验层面

测验层面的差异是指由于不同测验形式具有的特征不同而导致的差异, 比如 PBT 与 CBT 在作答设备、作答过程中是否允许检查并修改答案、测验过程中有无监督以及测验计时和选题方式等方面都具有不同的特征。具体来说:

(1)作答设备。在 PBT 中, 被试通常使用纸笔进行作答; 而在 CBT 中, 被试需要在显示屏上阅读题目, 并使用鼠标和键盘进行作答。屏幕大小、分辨率和刷新速度等都可能对被试在计算机上的作答产生影响。Ziefle(1998)对被试在 PBT 和两种屏幕分辨率(1664×1200 和 832×600)下 CBT 的阅读表现及感受到的疲劳程度进行比较, 结果发现: 被试在 PBT 中的表现显著好于两种分辨率下 CBT 的表现; 而且分辨率越高, 被试感受到的疲劳程度越轻。在屏幕大小方面, 其对 TME 的影响因人而异, 但总体来说, 更大的屏幕会增加文字的可读性, 从而提高测验表现(Bridgeman et al., 2003)。

(2)是否允许检查并修改答案。在 PBT 中, 被试可以不按题目的呈现顺序进行作答, 甚至可以随时对已作答题目进行检查并修改答案; 而有些 CBT (如计算机化自适应测验[Computerized Adaptive Testing, CAT])一般不允许被试返回检查并修改答案, 主要是因为考试机构担心提供修改机会会带来两个问题: ①“聪明”被试或“聪明”备考机构所指导的被试通过采用 Wainer 策略(Wainer, 1993)和 Kingsbury 策略(Wise et al., 1997)等作弊策略获得虚高的分数, 从而影响测验的公平性、公正性和准确性; ② 增加测验时间, 相应地增加测验费用。CAT 不提供修改功能也会给被试带来两方面的影响: ① 被试在 PBT 中惯用的作答策略不能用于 CAT, 会给他们带来焦虑和压力; ② 若被试完全有能力答对某道题目但是键入或点击失误了, 不允许修改会导致其能力被低估; 相反, 若被试没有能力答对某道题目但是猜对了, 不允许修改会导致其能力被高估(陈平, 丁树良, 2008; 高旭亮 等, 2016; 林喆 等, 2015)。不提供修改机会的 CAT 可能导致 TME 的产生。

(3)测验过程有无监督。一般情况下, PBT 的实施过程中往往有主试在场监督; 而对于部分 CBT (比如通过网络进行的在线测验)很有可能会在无人监督的情况下开展, 这也有可能导致 TME 的产生。Goldberg 和 Pedulla(2002)比较被试在 PBT、有监督 CBT 和无监督 CBT 的 GRE 分数, 结果表明: 被试在 PBT 和有监督 CBT 中的表现显著好于无监督 CBT。测验过程有无监督可能会对被试的作答动机产生影响, 从而影响其在测验中的表现。

(4)测验计时与选题方式。在 CBT 中, 计算机为更精细的考试流程设计提供了可能: ① 测验开发者可以将测验的计时设计为“以单道题目为单位”、“以测验模块为单位”或“以整个测验为单位”; ② 测验的组卷不再拘泥于固定试题, 而允许被试作答与自身能力匹配的题目(即 CAT)。虽然没有研究直接表明不同的测验计时设计会引起 TME, 但是相比于以单道题为单位的计时, 目前

主流的大型 CBT (如 PISA 和 NAEP)通常以一个测验模块为单位进行计时，且部分 CBT (如 GRE)允许被试选择偏好的时间呈现方式(即显示或不显示倒计时)。另外，相比于可能包含简单题的 PBT，CAT 中高能力水平被试的测验过程可能更“吃力”，因为总是作答与自身能力水平匹配的难题。为探究 CAT 匹配被试能力的选题策略是否会增加被试的测验焦虑程度进而引起 TME，Powers(1999)基于 GRE 的 PBT 和 CBT 样本进行回归分析，发现被试在两种测验形式下的焦虑与 GRE 分数之间的关系并无显著差异，而且自适应的选题策略并未加剧被试的测验焦虑。Fritts 和 Marszalek(2010)分析中学生的学业进度测验(measures of academic progress)结果后发现：在控制被试对考试的基线焦虑水平和对电脑使用的焦虑后，相比于 CAT，被试在 PBT 上表现出更高的焦虑水平。

2.2 题目层面

题目层面的差异来源于题目本身的属性，这些属性可能在不同测验形式下的表现不同，从而导致 TME 的产生。具体包括：

(1)题目呈现方式。呈现方式包括题目的字体、字号、粗细和颜色(Bernard et al., 2002; Bernard & Mills, 2000)、每一行的文字长度(Chaparro et al., 2002)、每一页中呈现的题目数量和行数(Duchnicky & Kolers, 1983)以及每一页中空白部分的面积大小(McMullin et al., 2002)等。由于 CBT 的形式多样且多借助现成软件或平台进行施测，很难保证所有题目都以相同方式呈现给被试，从而导致 TME 的产生。

(2)题目类型。题目类型可能会影响被试和题目间的交互方式，从而影响被试的作答表现(Kröhne & Martens, 2011)。题目类型主要包括两大类：选择题与建构题。对于选择题，特别是当题目较短时，不同测验形式的差异较小，较少检测出 TME(Buerger et al., 2016; Lynch, 2022)。而对于建构题，考生在 PBT 上的表现倾向于比 CBT 更好(Bennett et al., 2008)。这可能源于题目交互方式的复杂程度的变化，交互方式较复杂的题目更容易影响被试在 CBT 上的成绩(Kingston, 2008)。例如，当题目包含较长的文本或作答过程涉及使用鼠标、滚轮和下拉菜单等，题目的作答难度会增加(Poggio et al., 2005)。另外，Liu 等人(2016)对美国基础教育评价系统(PARCC)的数学建构题进行分析后发现：相对于 PBT 被试群体，CBT 被试群体在低年级(3~8 年级)题目上的表现更好；而对于高年级的建构题，结论则相反。这意味着题型在不同测验形式上的差异还可能源于题目所涉及的认知过程不同。Johnson 和 Green (2006)通过观察和访谈小学生后发现，约 1/3 的被试在作答不同测验形式下的题目时会采用不同的作答策略。而对于作文任务，研究认为被试在 CBT 上的表现优于 PBT，或两者没有显著差异(Lee, 2002; Lynch, 2022; Zhi & Huang, 2021)。Li (2006)让被试在作答学术英语任务(English for academic purposes)时进行出声思维，发现被试在 CBT 上展现出更高阶的思维能力，并且比 PBT 做出更多的修改。相比于关注单词水平的修改，被试在 CBT 上更多地进行句子和段落层次的完善和组织(Chan et al., 2018)。

2.3 被试层面

被试层面的差异来源于被试本身的属性，这些属性并非测验想要测量的特质，但是它们在不同测验形式上的差异可能会导致 TME 的产生。具体包括：

(1)人口学变量。性别、年龄、种族和社会经济地位等人口学变量并不直接导致 TME，而是通过影响与测验目的相关的被试能力等来间接导致 TME。比如，老年人可能由于使用计算机的熟练程度不如年轻人，因而在 CBT 上的表现更差(Chua et al., 1999)；但也有研究表明，年龄带来的差异并不像研究者预期的那样显著(Weigold et al., 2016)。Fouladi 等人(2002)发现不同测验形式间的结果存在较大差异，但在控制性别和种族的影响后，不同测验形式间的结果差异显著减小。

(2)计算机的熟练程度。对计算机使用越熟练，在 CBT 中的表现就越好(Jerrim et al., 2018; Pomplun, 2007)。一方面，对计算机越熟练，在作答时的操作就越快捷；另一方面，被试的学习过程和测试过程的形式相匹配时，他们的作答分数会更高，即存在一定的迁移适用加工过程(transfer appropriate processing; Clariana & Wallace, 2002)。但也有研究发现，使用计算机的熟练程度不会对被试在 CBT 上的结果产生影响(Jeong, 2012)。

(3)作答动机。与低利害测验相比，参加高利害测验的被试具有更高的作答动机，从而在 PBT 和 CBT 上有更相近的表现(Rowan, 2010)。有意思的是，也有研究发现：与 PBT 相比，被试对 CBT

普遍有更好的体验、更高的作答动机和自我效能感, 但在 CBT 上的得分却更低(Chua, 2012)。

2.4　评分者层面

评分者层面的差异本质上源于评分者内在认知加工的不同, 认知加工的不同可能使得评分者在不同测验形式下的评分结果有所差异, 从而导致TME的产生。也即, 评分者效应(rater effect; 韩建涛 等, 2019)也可能是 TME 的来源之一。测验中的客观题由于评分标准明确、客观, 所以其评分结果不易被评分者效应影响; 而对于主观题, 其评分结果则容易受到评分者主观因素的影响, 从而导致其在 PBT 和 CBT 中的评分结果存在差异。具体来说, 评分者在评定不同测验形式下的被试作答时, 主要受到被试作答呈现方式的影响(Hunsu, 2015), 其中手写版(handwritten)和打字版(typed or word-processed)的差异是研究关注的重点。Arnold 等人(1990)发现, 评分者倾向于对手写版作答采用更宽松的标准, 而对打字版更苛刻。这可能是因为手写作答在一定程度上具有更长的感知视觉效果, 并保留被试的修改痕迹, 而且有评分者认为手写版比打字版更有“力量” (Powers et al., 1994; Russell & Tao, 2004a)。另外, 为探讨不同测验形式给评分者带来的感知长度差异对测验结果的影响, 研究者对比单倍行距与双倍行距的作文评分, 发现长度的变化并没有消除 CBT 与 PBT 的得分差异(Russell & Tao, 2004b)。

需要注意的是, 来自评分者的影响通常与题型相互交织, 评分者对 CBT 与 PBT 的评分差异大多出现于建构题上。为区分两者的影响, 研究者将手写版的作答输入计算机, 让评分者对混合之后的打字版作答进行评分, 发现被试在 CBT 上的得分更高(Jin & Yan, 2017; Russell & Haney, 1997)。但也有对学术英语测试的研究发现, 控制评分者的严格程度和信度之后, 被试在 CBT 与 PBT 下的整体测验得分差异较小, 评分者仅在词汇量测试题中呈现出对手写版的偏好(Chan et al., 2018)。

表 1 对 TME 的来源进行总结, 并对 TME 的产生进行说明。

在实践中, 研究者往往需要在排除无关变量的影响后, 再探究测验形式对测验结果的影响。因此, 对 TME 的来源进行梳理有助于研究者在实验设计阶段对无关变量进行严格控制, 以减少无关变量的影响。比如, 在测验层面保证被试都能检查并修改答案, 且作答过程都在有人监督的情况下进行; 在题目层面保证所有题目在 PBT 和 CBT 上有相同的呈现效果; 在被试层面保证在两种测验形式上作答的被试的年龄和性别等方面一致。

3　TME 的检测

3.1　TME 的实验设计

TME 研究一般采用两类实验设计控制被试

表 1　TME 的来源和对 TME 产生的说明

TME 的来源		TME 产生的说明
测验层面		
	作答设备	PBT 使用纸笔, CBT 使用屏幕、鼠标和键盘
	是否允许检查并修改答案	PBT 允许检查修改答案, CBT 往往不允许
	测验过程有无监督	PBT 往往有监督, CBT 可能无监督
	测验计时与选题方式	CBT 的计时和选题方式更灵活, PBT 的更固定
题目层面		
	题目呈现方式	CBT 的多样形式导致很难与 PBT 有完全相同的题目呈现方式
	题目类型	题型交互方式的复杂程度影响 CBT 上的表现
被试层面		
	人口学变量	年龄和性别等通过影响其他变量间接导致 TME
	计算机的熟练程度	计算机熟练程度可能影响 CBT 上的成绩
	作答动机	PBT 和 CBT 上的作答动机不同导致得分差异
评分者层面		
	评分者效应	主观题易受评分者效应的影响

特征: 组间设计和组内设计(Buerger et al., 2016)。在 TME 的研究背景下, 组间设计中每名被试只接受 PBT 或者 CBT, 而在组内设计中每名被试先后接受这两种测验形式。TME 组间设计和组内设计如图 1 所示(共 *N* 名被试和 *I* 道题)。根据被试是否能够自由选择测验形式, 组间设计又被分为两类: (1)自由选择。即被试可以自由选择测验形式(Puhan et al., 2007); (2)随机分配。即研究者将被试随机分配给某种测验形式(Gu et al., 2021; Schwarz et al., 2003)。根据被试作答顺序是否固定, 组内设计也可以被分为两类: (1)固定顺序。即所有被试接受两种形式测验的顺序固定且一致(Jeong, 2012); (2)平衡顺序。即先将被试随机分成两组, 一组先接受测验形式 A (如 PBT), 一段时间后再接受测验形式 B (如 CBT), 另一组则与之相反, 即所谓的"AB-BA 设计" (Bodmann & Robinson, 2004; Kim et al., 2018; Seifert & Paleczek, 2022)。

组间设计和组内设计各有其适用范围。与前者比, 后者能有效避免由组间个体差异带来的无关变量干扰, 但也容易受到疲劳效应和练习效应的影响, 因此适用于样本量和题量都较少的情况, 更适用于练习效应较小的人格测验。而在组间设计中, 虽然组间个体差异难以避免、容易引入无关变量, 但是由于每名被试只接受一种测验形式, 实施起来更方便、快捷, 因而适用于样本量和题量都较多的情境, 更适用于能力测验。

为改进这两种设计的不足, 研究者将它们结合形成平衡不完全区组(Balanced Incomplete Block, BIB; Brunfaut et al., 2018)设计, 如表 2 所示。在 BIB 设计中, 原测验被分成多个平行题本, 相应地被试也被随机分成多个组, 这多个被试组理论上可被看作是相互平行的。表 2 中的"Test 1"和"Test 2"代表被试的作答顺序。每组被试作答两个题本, 并在题本序号和作答顺序上进行平衡, 从而减轻被试的疲劳效应。由于题本 A 和 B 理论上平行, 比较每组中两个题本间的作答就可以估计 TME。通过设计组 1 和组 4 以及组 2 和组 3 可以控制顺序效应、疲劳效应和学习效应。BIB 设计结合两种设计的优点, 因而在样本量大、题目较多的测评项目(如 PISA)中已经得到较为成熟的运用(OECD, 2014)。

通过实验设计, 可以有效控制组间被试特征的影响。但是即使控制组间差异, BIB 设计依旧无法完全避免组内个体差异(如年龄、计算机的使用和作答动机)的影响, 此时可以在测验过程中估计由个体特征造成的 TME。接下来介绍 TME 的检测方法。

3.2 TME 的检测方法

对 TME 进行检测就是对被试在 PBT 和 CBT 上的作答表现进行比较, 作答表现的比较可以分为两个层面: 观测变量层面和潜变量层面。在观测变量层面, 一般采用方差分析(Analysis of Variance, ANOVA)法进行比较。在潜变量层面, 一般通过检验测量不变性或参数不变性来检测 TME。在结构方程模型框架下, 测量不变性是指在测量被试的目标特质时, 观测变量和潜在特质间以及潜在特质之间的关系在待比较的各组之间或在不同情境下等同(白新文, 陈毅文, 2004); 而在项目反应理论(Item Response Theory, IRT)框架下, 参数不变性体现在题目参数和能力参数的不变性上(聂旭刚 等, 2018)。目前, 潜变量层面的 TME 检测方法主要包括多组验证性因子分析(Multigroup Confirmatory Factor Analysis, MCFA)法、题目功能差异(Differential Item Functioning, DIF)法和模式效应模型(Mode Effect Model, MEM)法。下面对这 4 种方法进行述评。

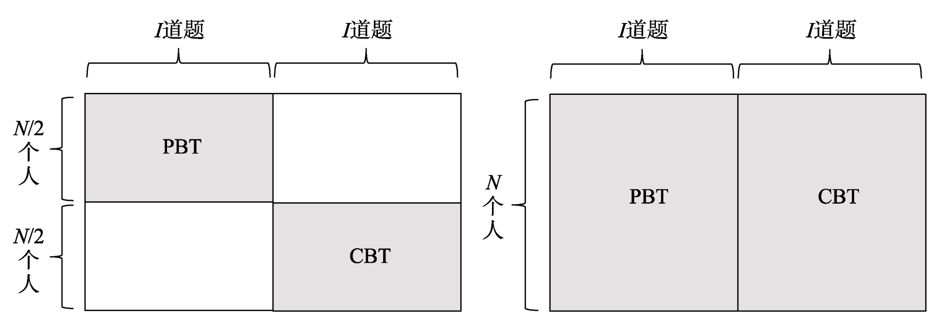


图 1 TME 组间设计(左)和组内设计(右)示意图

表 2　TME 研究中的 BIB 设计

组别	PBT		CBT	
	题本 A	题本 B	题本 A	题本 B
组 1	Test 1			Test 2
组 2		Test 1	Test 2	
组 3		Test 2	Test 1	
组 4	Test 2			Test 1

3.2.1　ANOVA 法

ANOVA 法首先计算两种测验形式下的作答指标(包括测验层面的总分以及题目层面的平均分、正确率和空缺率等), 然后根据实验设计是组内或组间设计, 采用被试内或被试间的 ANOVA 对这些作答指标进行比较。如果这些指标间存在显著差异, 则说明存在 TME 并且 TME 会对测验结果产生影响(Bodmann & Robinson, 2004; Goldberg et al., 2003; Khoshsima et al., 2017)。ANOVA 法可通过 SPSS 或 R 中的 TAM 包(Robitzsch et al., 2022)实现。

3.2.2　MCFA 法

MCFA 法采用多组比较的思想, 对两种测验形式下的结果进行测量不变性检验(Kim & Huynh, 2008)。测量不变性检验是通过比较一系列嵌套模型来实现, 具体表现在依次对以下不变性进行检验: (1)结构不变性(configural invariance)检验。即检验不同组之间的因子结构(即观测变量和潜变量间的关系)是否相同; (2)弱不变性(weak invariance)检验。若结构不变性得到满足, 则进一步检验不同组之间的因子载荷是否相等; (3)强不变性(strong invariance)检验。若弱不变性得到满足, 则进一步检验不同组之间的截距(潜变量预测观测变量时的截距)是否相同; (4)严格不变性(strict invariance)检验。若强不变性得到满足, 则检验不同组之间的残差方差是否相同。这 4 种检验对应的测量不变性水平具有层级嵌套关系, 只有低一级的不变性得到证实后, 进行高一级的不变性检验才有意义(蔡华俭 等, 2008)。如果测验在某一级水平的测量不变性上出现违反, 则说明该测验在对应水平上存在 TME, 通过这种方式可以对测验层面的 TME 进行检验。

为进一步寻找违反测量不变性的成因, 可以根据输出结果确定当前测量不变性水平下对模型拟合违反较大的题目。在放松该题目上的检验限制后, 若模型拟合显著变好, 则说明该题目的存在会对测量不变性产生影响, 可认为存在 TME。这样依次对所有题目进行检测, 即可找出所有具有 TME 的题目。此时, 测验满足部分(partial)弱不变性、部分强不变性或部分严格不变性。

已有研究几乎都得到结构不变性的结果, 这可能是因为一个用于施测的成熟测验往往具有较好的信效度, 所以在测验形式发生变化后因子结构并没有发生变化。大多数测验具有完全或部分弱不变性, 还有一些测验具有完全或部分强不变性, 但是极少有测验能够达到严格不变性(比如, Hox et al., 2015)。一般来说, 只要达到弱不变性或部分强不变性, 就说明不同测验形式下的结果可比。MCFA 法可通过 R 中的 lavaan 包(Rosseel, 2012)实现。

3.2.3　DIF 法

TME 和 DIF 都反映"由于某种因素的影响, 导致能力相同的被试在同一题目上具有不同的正确作答概率", 在 DIF 中这种因素是指被试来自不同群体, 而在 TME 中这种因素是指不同的测验形式。鉴于两者的相似性, 不少研究者将检测 DIF 的方法用于对 TME 的检测(Chan et al., 2004; Keng et al., 2008; Puhan et al., 2007; Schwarz et al., 2003), 此时作答 CBT 的被试组可看作是目标组(focus group), 作答 PBT 的被试组可看作是参照组(reference group)。

常见的 DIF 检测方法主要有两类: 一类是基于 IRT 的方法(即将潜在特质作为匹配变量), 包括 IRT 似然比检验法(IRT Likelihood Ratio, IRT-LR)、测验与题目功能差异法(Differential Functioning of Items and Test, DFIT)以及同时题目偏差检验法(Simultaneous Item Bias Test, SIBTEST; Shealy & Stout, 1993)等; 另一类是非 IRT 的方法(即直接将测验总分作为匹配变量), 包括 Mantel-Haenszel 法、标准化法(Standardization, STND)和逻辑斯蒂克回归法(Logistic Regression, LRDIF)等。其中, Mantel-Haenszel、SIBTEST、IRT-LR 和 DFIT 法都已被用于检测 TME(Claudia et al., 1999; Puhan et al., 2007; Terluin et al., 2018)。值得注意的是, 只有 DFIT 法可以同时对测验和题目层面的 DIF 进行检测, 其他方法只能对单个题目的 DIF 进行检测(Raju et al., 1995)。

以 SIBTEST 法为例, 简要介绍检测 TME 的

步骤: (1)将所有题目分为匹配子测验和待测子测验。匹配子测验由不存在 TME 的题目组成, 因此可将被试在匹配子测验上的分数作为其能力估计值; (2)对目标组和参照组在匹配子测验和待测子测验中的作答结果进行评价, 并基于匹配子测验上的分数将能力相同但组别不同的被试进行匹配。SIBTEST 假定在匹配子测验中分数相同的被试具有相同能力, 所以组别不同的匹配被试在待测子测验上的分数差异就是 TME 的值; (3)对 TME 的值进行显著性检验, 从而确定题目是否有 TME(蔡晓芬, 2014; 汤楚, 2016)。DIF 法可通过 R 中的 mirt 包(Chalmers, 2012)实现。

3.2.4 MEM 法

von Davier 等人(2019)提出可以通过在两参数逻辑斯蒂克模型(Two-Parameter Logistic Model, 2PLM)中加入量化的 TME 参数从而形成 MEM, 然后在估计题目参数和能力参数的同时也对 TME 参数进行估计。MEM 包含三个子模型, 每个子模型都有不同的模型假设。

MEM 中的模型 1 又被称为一般 MEM(general MEM)。它假设 TME 只与测验形式有关, 在测验形式发生变化后, 所有题目的难度都发生相同的改变。模型 1 定义 TME 参数为 δ_m(m 代表测验形式), 公式如下:

$$P(x=1\mid\theta,\alpha_i,\beta_i,\delta_m)=\frac{\exp(\alpha_i\theta-\beta_i-1_{\{i>I\}}\delta_m)}{1+\exp(\alpha_i\theta-\beta_i-1_{\{i>I\}}\delta_m)} \quad (1)$$

其中 α_i 和 β_i 分别为第 i 题的斜率参数和截距参数, θ 为能力参数, I 表示一种测验形式的测验的题目数量。$1_{\{i>I\}}$ 是指示函数, 当 $i \leqslant I$ 时, $1_{\{i>I\}}$ 为 0, 代表原本的测验形式(如 PBT); 令 $i'=i+I$, 即 $I<i'\leqslant 2I$ 时, $1_{\{i'>I\}}$ 为 1, 代表新的测验形式(如 CBT)。此时, 第 i 题和第 i' 题是同一道题目, 但测验形式不同。为使作答 PBT 和 CBT 的被试在同一题目上的正确作答概率一致, 模型假设 $\alpha_i=\alpha_{i'}$ 和 $\beta_i=\beta_{i'}+\delta_m$。当 $\delta_m=0$ 时, 说明测验在 PBT 和 CBT 间不存在显著差异, 即测验无 TME; 当 $\delta_m>0$ 时, 则有 $\beta_i>\beta_{i'}$, 说明测验在 PBT 上的难度大于 CBT; 当 $\delta_m<0$ 时, 说明测验在 PBT 上的难度小于 CBT[1]。

MEM 中的模型 2 假设测验形式和题目之间存在交互作用, 也即在测验形式发生变化后, 测验中有的题目可能会变得更难, 有的题目会变得更简单。因此, 模型 2 也被称为题目特异性的 MEM (item-specific MEM), 公式如下:

$$P(x=1\mid\theta,\alpha_i,\beta_i,\delta_{mi})=\frac{\exp(\alpha_i\theta-\beta_i-1_{\{i>I\}}\delta_{mi})}{1+\exp(\alpha_i\theta-\beta_i-1_{\{i>I\}}\delta_{mi})} \quad (2)$$

与模型 1 类似, 模型 2 中的前 I 道题对应 PBT、后 I 道题对应 CBT。两种测验形式上的题目一一对应, 因此也有 $\alpha_i=\alpha_{i'}$ 和 $\beta_i=\beta_{i'}+\delta_{mi}$, δ_{mi} 为第 i 题的 TME 参数。当 $\delta_{mi}=0$ 时, 说明第 i 题不存在 TME; 当 $\delta_{mi}>0$ 时, 说明在第 i 题上 PBT 的难度大于 CBT; 当 $\delta_{mi}<0$ 时, 说明在第 i 题上 PBT 的难度小于 CBT。

MEM 中的模型 3 假设测验形式和被试之间存在交互作用, 即在测验形式转化后, 对于有的被试来说题目变得更难, 对于有的被试来说题目变得更简单。模型 3 也被称为个体特异性的 MEM (person-specific MEM), 公式如下:

$$P(x=1\mid\theta,\alpha_i,\beta_i,a_{mi},\vartheta)=\frac{\exp(\alpha_i\theta-\beta_i-1_{\{i>I\}}a_{mi}\vartheta)}{1+\exp(\alpha_i\theta-\beta_i-1_{\{i>I\}}a_{mi}\vartheta)} \quad (3)$$

其中 a_{mi} 是模式斜率, 它具有题目特异性, 反映个体特征对 TME 的影响在不同题目上不同。ϑ 代表被试的额外能力(如使用计算机的能力), 它与 TME 有关, 但与被试的与测验目的有关的能力不相关, 即 $\mathrm{cov}(\theta,\vartheta)=0$。如果 $a_{mi}=0$, 说明不存在 TME; 如果 a_{mi} 显著不等于 0, 则存在 TME。模型 3 与前两个模型的最大区别在于: 模型从单维 IRT 模型变成多维 IRT 模型, 因此在模型识别和参数估计上都更复杂。

MEM 法的以上三个子模型分别假设三种不同的情况。在使用这种方法检测 TME 时, 通常的做法是使用 AIC 和 BIC 等模型拟合指标比较三个模型和数据的拟合程度, 拟合越好说明数据更接近对应模型的假设, 从而可以探究 TME 是具有一般性、题目特异性还是个体特异性(von Davier et al., 2019)。模型拟合的同时也对题目参数、能力参数和 TME 参数进行估计, 进而找出具有 TME 的题目并对其进行调整。另外, 模型 1 和 2 具有嵌套关系, 模型 3 与模型 1 和 2 没有嵌套关系。如果简单模型和复杂模型的拟合不存在显著差异, 则

---

[1]公式(1)中若没有 $1_{\{i>I\}}\delta_m$ 部分, 即对应 2 PLM。在 2 PLM 中, β_i 与题目难度 b_i 成正比关系, 即 $\beta_i=b_i\times\alpha_i$, 其中 $\alpha_i=1.702\times a_i$(a_i 是题目区分度)。

选择性价比更高的简单模型。MEM 法可通过 mdltm 软件(von Davier, 2005)实现。

MEM 法的三个子模型还可以从 TME 来源的角度进行理解。模型 1 假设 TME 只与测验形式有关, 说明此时 TME 的来源只包括测验层面的差异, 如计算机的硬件设施和是否允许检查并修改答案等。模型 2 假设 TME 具有题目特异性, 说明此时 TME 会受到题目层面差异的影响, 如题目类型和题目的呈现方式等。这种情况在能力测验中较为常见, 特别是包含多种题型的考试中, 不同题目受到测验形式的影响也不同, 从而导致题目特异性的 TME。模型 3 假设 TME 具有个体特异性, 说明此时 TME 会受到被试层面差异的影响, 如年龄、性别、计算机的熟练程度和作答动机等。这种情况可能出现在个体差异较大的时候, 即使通过实验设计进行控制, 也没法完全避免个体差异的影响, 从而导致个体特异性的 TME。

为促进 TME 检测方法的应用, 本文在附录部分呈现能实现 ANOVA、MCFA 和 DIF 方法的 R 代码示例, 并以组间设计为例给出检验题目层面 TME 的简要流程。

3.2.5　TME 检测方法的比较

表 3 对上述 4 种 TME 检测方法的优缺点、适用范围和实现方法进行了总结。

ANOVA 法通过“计算 PBT 和 CBT 上的作答指标, 再比较两者间的差异”来检测 TME, 优点在于方便快捷、计算简单, 适合对测验层面的 TME 进行初步检测; 不足在于检验力较低, 而且只能对观测指标进行比较。MCFA 法通过验证测量不变性来对 TME 进行检测。与 ANOVA 法类似, MCFA 法更适合对测验层面的 TME 进行检测, 可以探究观测变量与潜在特质间以及潜在特质间的关系; 不足在于对题目层面 TME 进行检测的过程繁琐、不易操作。

DIF 法利用 DIF 和 TME 在概念和检测方法上的共通性, 采用 DIF 检测方法对 TME 进行检测。DIF 法的优点体现在两方面: 一是能对测验中具有 TME 的题目进行准确识别; 二是包含的方法非常多样, 在实践中可以灵活选择。MEM 法通过建立包含 TME 参数的 IRT 模型, 直接对 TME 的值进行估计。与前三种方法相比, MEM 法具有两方面的优点: 一是能对 TME 的大小进行直接估计; 二是能在一定程度上探究 TME 的来源, 从而更好地对 TME 进行解释和控制; 缺点是模型较为复杂(特别是模型 3), 可能会面临模型识别和参数估计等方面的挑战。

4　测验模式效应的研究结果

在过去 30 多年里, 已经有超过 300 项研究对 PBT 和 CBT 的测验结果(包括成就测验、人格与态度测验和职业兴趣测验等领域的结果)进行比较(Duchnicky & Kolers, 1983; Kulik et al., 1980), 但并没有得到一致的结论。很多研究者发现, 同一测验在 CBT 上的难度要普遍高于 PBT, 导致被试在 PBT 上的表现显著好于在 CBT 上的表现(比如, Backes & Cowan, 2019; Beatty et al., 2022; Lee et al., 1986; Jeong, 2012)。然而也有一些研究得出相反的结论, 即被试在 CBT 上的表现要好于在 PBT 上的表现(比如, Brunfaut et al., 2018; Russell & Plati, 2002)。还有不少研究发现, 被试在不同测验形式上的作答结果没有显著差异(Blumenthal & Blumenthal, 2020; Hamhuis et al., 2020; Khoshsima & Toroujeni, 2017; Paleczek et al., 2021; Porion et al., 2016; Prisacari & Danielson, 2017a, 2017b)。

出现这样的结果可能与研究发表的年代有关。随着研究发表年代的递进, 被试在 PBT 和

表 3　四种 TME 检测方法的总结

	优点	缺点	适用范围	实现方式
ANOVA	方便快捷, 适用范围广	检验力较低	对 TME 进行初步检测	SPSS 或 TAM 包
MCFA	可探究潜变量和观测变量间以及潜变量间的关系	对题目层面的 TME 检测过程较为繁琐	人格和社会心理领域内的测验	lavaan 包
DIF	检验力高, 包含方法多样, 可灵活选择	各种 DIF 方法的自身不足	教育测量领域内的成就测验	mirt 包
MEM	检验力高, 可在一定程度上了解 TME 的来源	模型较为复杂, 可能出现模型识别等问题		mdltm 软件

CBT 上的作答表现也发生变化。在 21 世纪之前，计算机还没有得到普及，相应地人们对计算机的使用不太熟练，因此会出现 PBT 得分显著高于 CBT 的结果。随着计算机的逐渐普及，人们使用计算机的能力也得到提高，再加上对计算机有着较强的兴趣和作答动机，因此出现更多在 CBT 上得分更高的情况。

对于没有检测出 TME 的研究，则可能有以下几点原因：(1)部分测验题目(如多选题)的稳定性较好，不易产生 TME; (2)随着题型越来越多样化，可能会出现“在同一测验中，部分题目对 PBT 更有利，而另一些题目对 CBT 更有利”的情况。如果只对测验层面的 TME 进行检测，则可能出现效应上的抵消; (3)在“测验本身结构较好、实验设计较完善且对 TME 来源控制较好”的前提下，测验层面不存在较大的 TME。若研究者采用检验力较低的 ANOVA 和 MCFA 法，则容易出现 TME 检测不显著的情况。

因此，很多研究在对测验层面的 TME 进行检测后，还会对题目层面的 TME 进行检测(Keng et al., 2008; Puhan et al., 2007; OECD, 2017)。通过综合测验和题目层面的检测结果，可以为测验在 PBT 和 CBT 上的可比性提供依据，也可以更细致地探究 TME 的来源，从而为题目的修订提出建议。

5 讨论与展望

目前随着计算机和网络的广泛运用，TME 已经成为大型测验电子化进程中不容忽视的问题。PISA、NAEP 和 TIMSS 等大规模测评项目都在经历着从 PBT 到 CBT 的变化。在进行测验形式的转变之前，采用严密的实验设计和精确的检测方法对测验中可能存在的 TME 进行检测，是保证 PBT 和 CBT 上作答结果具有可比性的重要途径，也是对测验公平的保障。

通过前面的梳理，可以看到尽管 TME 的研究已经较为成熟，但是也还存在一些问题：首先，TME 的来源比较复杂，使得影响 TME 的因素繁多。而且对于同一因素，还可能会在不同人群中出现巨大差异。比如 CBT 中的交互方式，年轻人会适应键盘和鼠标的输入方式，而中老年人可能会非常不适应。这使得研究者几乎无法预测和控制影响因素，不利于对 TME 进行深入的分析与解释。其次，缺少对 TME 检测方法的系统比较。尽管 4 种 TME 检测方法各有优势，有时也可以同时使用以达到更好的效果，但是还没有研究对它们的检测效果进行全面比较。最后，不同 TME 研究中的结果难以进行比较。如前所述，TME 的研究结果受 TME 的来源、实验设计和检测方法等多方面的影响，因此有研究者使用元分析方法对 TME 研究的影响因素进行探究，然而结果不尽相同(Wang et al., 2007, 2008)。这可能是因为元分析本身存在“苹果与桔子之争”问题，即很多研究者认为方法不同的研究不能进行直接比较。

综上，TME 今后的研究方向包括但不限于以下几个方面:

5.1 提升 MEM 方法的解释性与适用性

第三部分提到，可以从 TME 来源的角度理解 MEM。但是, MEM 只能在一定程度上帮助研究者锁定 TME 的来源范围，无法对 TME 的来源做出解释。因此，可以借助“IRT 模型能够增减参数”的优势，在现有 MEM 中加入与 TME 来源相关的因素，从而直接在模型中对 TME 进行解释。比如，模型 1 假设 TME 只与测验形式有关, TME 的来源可能是作答过程有无监督等测验层面的特征。为进一步对这些因素进行解释，可以建立关于 TME 参数和测验层面特征的回归方程，以探究不同特征的权重以及不同特征对 TME 产生的贡献大小。在模型 2 和 3 中，也可以建立类似的回归方程对 TME 的来源进行解释。

另外，还可以使用广义模型对 TME 进行解释。陈冠宇和陈平(2019)基于广义线性混合模型和非线性混合模型的视角全面探讨解释性 IRT 模型(Explanatory IRT Model, EIRTM)。EIRTM 是一个综合性的解释框架，它通过在 IRT 模型中加入预测变量来对被试和题目间的关系进行刻画，进而解释相关变量的影响。具体地讲，他们在 EIRTM 的框架下，从固定效应和随机效应的角度对 TME 进行解释。未来研究也可以在 EIRTM 这一更加灵活、更加广义的框架下对混合 MEM 进行进一步界定。

再者，已有的 MEM 方法主要基于 IRT 模型(即 2PLM)。而认知诊断测验(Cognitive Diagnostic Testing, CDT)由于能够反馈学生对特定知识属性的掌握情况、能够剖析心理量表的潜在结构(de La Torre & Douglas, 2004)，正日益受到测量研究者和实践者的青睐。未来研究可进一步开发适用于

CDT 的 MEM 方法, 比如借助广义多策略认知诊断模型(Ma & Guo, 2019)分析 CBT 与 PBT 下的被试作答策略差异, 以了解不同测验形式下的认知加工过程变化。

5.2 拓展 TME 研究中测验形式的范围

目前大多数 TME 研究都聚焦于 PBT 和 CBT 之间的比较, 然而 TME 还可能出现在 PBT 和其他测验形式之间, 包括手机测验(mobile-based assessment)和电话或面对面访谈(phone or face-to-face interview)等测验形式(Chan et al., 2004; Magnus et al., 2016)。Kim 和 Walker(2021)还研究在考试中心参加测验和使用远程监考在家参加测验之间的 TME。随着测验形式的不断发展, 更多新型测验形式不断涌现, 比如基于游戏的测验(game-based assessment)、基于虚拟现实(virtual reality)和增强现实(augmented reality)等智能穿戴设备的测验等。对这些形式的测验进行 TME 研究也值得未来研究者重视。

5.3 将 TME 研究成果应用于我国大规模教育测评项目

在 PISA 2014 年的现场实验研究(field trial study)中, 研究者在参与测试的学校中随机选取学生参加 PBT 和 CBT, 并通过多种方法对 TME 进行检测, 证实数学、阅读和科学等认知测验在 PBT 和 CBT 上的结果具有可比性, 从而为测验形式的转变提供理论依据(OECD, 2016)。随后在 2015 年的正式测验中, 全球参与测试的 74 个国家(地区)中的 58 个国家(地区)全面使用 CBT 进行测验(OECD, 2017)。

而在我国的一些大规模教育测评项目中, 学科测验仍采用 PBT 的形式。这主要是因为我国各地的信息化水平程度不同、计算机或网络机房的配备程度不同, 导致少部分地区尚无条件使用 CBT。通过对 TME 进行深入研究, 可在一定程度上解决这一问题: (1)若测验中不存在显著影响测验结果的 TME, 则说明该测验在 PBT 和 CBT 上的结果具有测量等价性, 即可以在不同地区使用不同测验形式; (2)若测验中存在具有 TME 的题目, 则可以对其进行修订和改进, 增强它们在不同情境中的稳定性。

需要注意的是: 对于部分需要人工评分的建构题, 仍需尽量避免评分者对被试作答呈现方式感知差异所带来的影响。比如: (1)考虑将手写作答输入计算机, 能较有效地控制来自评分者层面的影响; (2)通过改良对评分者的训练规则来降低手写版和打字版的评分差异(Powers et al., 1994)。另外, 随着自动评分技术的发展(Ramesh & Sanampudi, 2022 ; Zhang et al., 2020), 测验或将迎来全计算机化模式, 届时评分者对 TME 的影响将主要集中在机器评分的算法层面。

参考文献

白新文, 陈毅文. (2004). 测量等价性的概念及其判定条件. *心理科学进展, 12*(2), 231−239.

蔡华俭, 林永佳, 伍秋萍, 严乐, 黄玄凤. (2008). 网络测验和纸笔测验的测量不变性研究——以生活满意度量表为例. *心理学报, 40*(2), 228−239.

蔡晓芬. (2014). *SP 程序和 DFTD 策略应用于 IRT 取向下 DIF 检测方法的效应比较* (硕士学位论文). 江西师范大学, 南昌.

陈冠宇, 陈平. (2019). 解释性项目反应理论模型: 理论与应用. *心理科学进展, 27*(5), 937−950.

陈平, 丁树良. (2008). 允许检查并修改答案的计算机化自适应测验. *心理学报, 40*(6), 737−747.

高旭亮, 涂冬波, 王芳, 张龙, 李雪莹. (2016). 可修改答案的计算机化自适应测验的方法. *心理科学进展, 24*(4), 654−664.

韩建涛, 刘文令, 庞维国. (2019). 创造力测评中的评分者效应. *心理科学进展, 27*(1), 171−180.

林喆, 陈平, 辛涛. (2015). 允许 CAT 题目检查的区块题目袋方法. *心理学报, 47*(9), 1188−1198.

聂旭刚, 陈平, 张缨斌, 何引红. (2018). 题目位置效应的概念及检测. *心理科学进展, 26*(2), 368−380.

檀慧玲, 李文燕, 万兴睿. (2018). 国际教育评价项目合作问题解决能力测评: 指标框架、评价标准及技术分析. *电化教育研究, 39*(9), 123−128.

汤楚. (2016). *短测验项目功能差异检测方法的比较研究* (硕士学位论文). 江西师范大学, 南昌.

Arnold, V., Legas, J., Obler, S., Pacheco, M. A., Russell, C., & Umbdenstock, L. (1990). *Do students get higher scoreson their word-processed paper? A study of bias in scoring hand-written vs. word-processed papers*. Retrieved March 7, 2023, from https://files.eric.ed.gov/fulltext/ED345818.pdf.

Backes, B., & Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Education Review, 68*, 89−103.

Beatty, A. E., Esco, A., Curtiss, A. B. C., & Ballen, C. J. (2022). Students who prefer face-to-face tests outperform their online peers in organic chemistry. *Chemistry Education Research and Practice, 23*, 464−474.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *The Journal of Technology, Learning, and Assessment, 6*(9), 1−39.

Bernard, M., Fernandez, M., Hull, S., & Chaparro, B. S. (2003). The effects of line length on children and adults' perceived and actual online reading performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 47*(11), 1375−1379.

Bernard, M., Lida, B., Riley, S., Hackler, T., & Janzen, K. (2002). A comparison of popular online fonts: Which size and type is best. *Usability News, 4*(1), 1−8.

Bernard, M., & Mills, M. (2000). So, what size and type of font should I use on my website? *Usability News, 2*(2), 1−5.

Blumenthal, S., & Blumenthal, Y. (2020). Tablet or paper and pen? Examining mode effects on German elementary school students' computation skills with curriculum-based measurements. *International Journal of Educational Methodology, 6*(4), 669−680.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research, 31*(1), 51−60.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16*(3), 191−205.

Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing, 36*, 3−18.

Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The transition to computer-based testing in large-scale assessments: Investigating (partial) measurement invariance between modes. *Psychological Test and Assessment Modeling, 58*, 597−616.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1−29.

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) scale: An item response theory analysis. *Medical Care, 42*(3), 281−289.

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing, 36*, 32−48.

Chua, S. L., Chen, D.-T., & Wong, A. F. L. (1999). Computer anxiety and its correlates: A meta-analysis. *Computers in Human Behavior, 15*(5), 609−623.

Chua, Y. P. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavior, 28*(5), 1580−1586.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593−602.

Claudia, P. F., Oshima, T. C., & Nambury, S. R. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*(4), 309−326.

de La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333−353.

Duchnicky, R. L., & Kolers, P. A. (1983). Readability of text scrolled on visual display terminals as a function of window size. *Human Factors, 25*(6), 683−692.

Feskens, R., Fox, J.-P., & Zwitser, R. (2019). Differential item functioning in PISA due to mode effects. In B. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 231−247). Cham, Switzerland: Springer.

Fouladi, R. T., McCarthy, C. J., & Moller, N. (2002). Paper-and-pencil or online? Evaluating mode effects on measures of emotional functioning and attachment. *Assessment, 9*(2), 204−215.

Fritts, B. E., & Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education, 13*, 441−458.

Goldberg, A., Russell, M. & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *The Journal of Technology, Learning, and Assessment, 2*(1), 1−52.

Goldberg, A. L., & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement, 62*(6), 1053−1067.

Gu, L., Ling, G. M., Liu, O. L., Yang, Z. T., Li, G. R., Kardanova, E., & Loyalka, P. (2021). Examining mode effects for an adapted Chinese critical thinking assessment. *Assessment & Evaluation in Higher Education, 46*(6), 879−893.

Hamhuis, E., Glas, C., & Meelissen, M. (2020). Tablet assessment in primary education: Are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students? *British Journal of Educational Technology, 51*(6), 2340−2358.

Hox, J. J., De Leeuw, E. D., & Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology, 6*, Article 87.

Hunsu, N. J. (2015). Issues in transitioning from the traditional blue-book to computer-based writing assessment. *Computers and Composition, 35*, 41−51.

Jeong, H. (2012). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology*, *33*(4), 410−422.

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice, 23*(4), 495−518.

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*(4), 476−493.

Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly, 14*(2), 101−119.

Johnson, M., & Green, S. (2006). On-Line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment, 4*(5), 1−35.

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education, 21*(3), 207−226.

Khoshsima, H., Hosseini, M., & Toroujeni, S. M. H. (2017). Cross-mode comparability of computer-based testing (CBT) versus paper-pencil based testing (PPT): An investigation of testing administration mode among Iranian intermediate EFL learners. *English Language Teaching, 10*(2), 23−32.

Khoshsima, H., & Toroujeni, S. M. H. (2017). Comparability of computer-based testing and paper-based testing: Testing mode effect, testing mode order, computer attitudes and testing mode preference. *International Journal of Computer, 24*, 80−99.

Kim, D., & Huynh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement, 68*(4), 554−570.

Kim, S., & Walker, M. (2021). *Assessing mode effects of at-home testing without a randomized trial* (ETS Research Reprot Series, No. 21-10). New Jersey, NJ: Educational Testing Service.

Kim, Y. J., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2018). Straightlining: Overview of measurement, comparison of indicators, and effects in mail-web mixed-mode surveys. *Social Science Computer Review, 37*(2), 214−233.

Kingston, N. M. (2008). Comparability of computer-and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*(1), 22−37.

Kline, R. (2013). Assessing statistical aspects of test fairness with structural equation modelling. *Educational Research and Evaluation: Fairness Issue in Educational Assessment, 19*(2-3), 204−222.

Kroehne, U., Gnambs, T., & Goldhammer, F. (2019). Disentangling setting and mode effects for online competence assessment. In H. P. Blossfeld & H. G. Roβbach (2nd Eds.), *Education as a lifelong process* (pp. 171−193). Wiesbaden, Germany: Springer VS.

Kröhne, U., & Martens, T. (2011). 11 Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, 14*, 169−186.

Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. (1980). Effectiveness of computer-based college teaching: A meta-analysis of findings. *Review of Educational Research, 50*(4), 525−544.

Lee, J. A., Moreno, K. E., & Sympson, J. B. (1986). The effects of mode of test administration on test performance. *Educational and Psychological Measurement, 46*(2), 467−474.

Lee, Y.-J. (2002). A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing, 8*, 135−157.

Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing, 11*(1), 5−21.

Liu, J., Brown, T., Chen, J., Ali, U., Hou, L., & Costanzo, K. (2016). *Mode comparability study based on spring 2015 operational test data*. Retrieved March 6, 2023, from https://files.eric.ed.gov/fulltext/ED599049.pdf.

Lynch, S. (2022). Adapting paper-based tests for computer administration: Lessons learned from 30 years of mode effects studies in education. *Practical Assessment, Research, and Evaluation, 27*, Article 22.

Ma, W., & Guo, W. (2019). Cognitive diagnosis models for multiple strategies. *British Journal of Mathematical and Statistical Psychology, 72*(2), 370−392.

Magnus, B. E., Liu, Y., He, J., Quinn, H., Thissen, D., Gross, H. E., & Reeve, B. B. (2016). Mode effects between computer self-administration and telephone interviewer-administration of the PROMIS(®) pediatric measures, self-and proxy report. *Quality of Life Research, 25*(7), 1655−1665.

McMullin, J., Varnhagen, C., Heng, P., & Apedoe, X. (2002). Effects of surrounding information and line length on text comprehension from the web. *Canadian Journal of Learning and Technology, 28*, 19−29.

OECD. (2014). *PISA 2015 Field Trial Goals, Assessment Design, and Analysis Plan for Cognitive Assessment*. PISA, OECD Publishing, Paris.

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. PISA, OECD Publishing, Paris.

OECD. (2017). *PISA 2015 technical report.* PISA, OECD Publishing, Paris.

Paleczek, L., Seifert, S., & Schöfl, M. (2021). Comparing digital to print assessment of receptive vocabulary with GraWo-KiGa in Austrian kindergarten. *British Journal of Educational Technology, 52*(6), 2145−2161.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*(6), 1−31.

Pomplun, M. (2007). A bifactor analysis for a mode-of-administration effect. *Applied Measurement in Education, 20*, 137−152.

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment, 11*(2), 127−143.

Porion, A., Aparicio, X., Megalakaki, O., Robert, A., & Baccino, T. (2016). The impact of paper-based versus computerized presentation on text comprehension and

memorization. *Computers in Human Behavior, 54*, 569−576.
Powers, D. E. (1999). *Test anxiety and test performance: Comparing paper-based and computer-adaptive versions of the GRE general test* (ETS Research Report Series, No. 99-15). Princeton, NJ: Educational Testing Service.
Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). They think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*(3), 220−233.
Prisacari, A. A., & Danielson, J. (2017a). Rethinking testing mode: Should I offer my next chemistry test on paper or computer? *Computers & Education, 106*, 1−12.
Prisacari, A. A., & Danielson, J. (2017b). Computer-based versus paper-based testing: Investigating testing mode with cognitive load and scratch paper use. *Computers in Human Behavior, 77*, 1−10.
Puhan, G., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning, and Assessment, 6*(3), 1−21.
Raju, N. S., van der Linden, W., & Fleer, P. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*(4), 353−368.
Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495−2527.
Robitzsch, A., Kiefer, T., & Wu, M. (2022). *Test Analysis Modules* (TAM). R package. Retrieved April 26, 2023, from https://cran.r-project.org/web/packages/TAM/TAM.pdf.
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1−36.
Rowan, B. (2010). *Comparability of paper-and-pencil and computer-based cognitive and non-cognitive measures in a low-stakes testing environment* (Unpublished doctorial dissertation). James Madison University, Harrisonburg.
Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*(3), 1−20.
Russell, M., & Plati, T. (2002). Does it matter with what I write? Comparing performance on paper, computer and portable writing devices. *Current Issues in Education, 5*(4), 1−15.
Russell, M, & Tao, W. (2004a). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research, and Evaluation, 9*, Article 1.
Russell, M., & Tao, W. (2004b). The influence of computer-print on rater scores. *Practical Assessment, Research, and Evaluation, 9*, Article 10.
Schwarz, R. D., Rich, C., & Podrabsky, T. (2003, April). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
Seifert, S., & Paleczek, L. (2022). Comparing tablet and print mode of a German reading comprehension test in grade 3: Influence of test order, gender and language. *International Journal of Educational Research, 113*, 1−13.
Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159−194.
Terluin, B., Brouwers, E. P. M., Marchand, M. A. G., & de Vet, H. C. W. (2018). Assessing the equivalence of web-based and paper-and-pencil questionnaires using differential item and test functioning (DIF and DTF) analysis: A case of the Four-Dimensional Symptom Questionnaire (4DSQ). *Quality of Life Research*, *27*(5), 1191−1200.
von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report Series, No. 05-16). Princeton, NJ: Educational Testing Service.
von Davier, M., Khorramdel, L., He, Q. W., Shin, H. J., & Chen, H. W. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics, 44*(6), 671−705.
Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*, 15−20.
Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219−238.
Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5−24.
Weigold, A., Weigold, I. K., Drakeford, N. M., Dykema, S. A., & Smith, C. A. (2016). Equivalence of paper-and-pencil and computerized self-report surveys in older adults. *Computers in Human Behavior, 54*, 407−413.
Wise, S. L., Freeman, S. A., Finney, S. J., Enders, C. K., & Severance, D. D. (1997, March). *The accuracy of examinee judgments of relative item difficutlty: Implications for computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.
Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating text generation with BERT*. arXiv preprint arXiv:1904.09675.
Zhi, M., & Huang, B. (2021). Investigating the authenticity of computer-and paper-based ESL writing tests. *Assessing Writing, 50*, Article 100548.
Ziefle, M. (1998). Effects of display resolution on visual performance. *Human Factors, 40*(4), 554−568.

Test mode effect: Sources, detection, and applications

CHEN Ping, DAI Yi, HUANG Yingshi

(*Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing* 100875, *China*)

Abstract: Test mode effect (TME) refers to the difference in test function caused by the administration of the same test in different test modes. The existence of TME will have an impact on test fairness, selection criteria and test equating, so it is of great significance to accurately detect and interpret TME. By systematically sorting out the source, detection (including the experimental design and detection methods) and research results of TME, the methodology of TME research is comprehensively demonstrated. Further interpretation of the TME model, expansion of the test modes in TME research, and application of TME research results to large-scale educational assessment programs in China, are important future development directions in the field of TME.

Keywords: test mode effect, test fairness, measurement invariance, computer-based testing

附录:

为促进 TME 检测方法的应用, 以下呈现能实现 ANOVA、MCFA 和 DIF 方法的 R 代码示例。由于实现 MEM 方法的 mdltm 软件不是开源软件且研究者在技术报告中并未提供详细的参数估计方法, 因此未囊括在本例中。接下来以组间设计为例, 给出检验题目层面 TME 的简要流程。

附表 1　基于 R 软件的 ANOVA、MCFA 和 DIF 方法代码示例

检验方法	代码示例

ANOVA	目的: 比较每一题在 PBT 和 CBT 上的平均分
	# 加载所需程序包 -------
	library(TAM)
	# 数据准备 ----------------
	# 1 = PBT, 0 = CBT
	# nperson 为被试量(即图 1 中 *N*)
	# nitem 为题目数(即图 1 中 *I*)
	# response_raw 包含两种测验形式下的所有作答, 是一个[nperson, nitem]的矩阵
	# TMEbetween 用于储存每道题在不同测验形式下的显著性结果
	
	# 创建数据框, 包含测验模式标签"mode"与相应的作答数据
	response_b <- data.frame(mode = c(rep(1, nperson/2), rep(0, nperson/2)),
	response_raw)
	
	# 数据分析 ----------------
	# 创建空矩阵用于结果存储
	TMEbetween <- matrix(data = NA, nrow = nitem, ncol = 1)
	for (j in 1:nitem){
	# 对每一题比较两种测验模式下的得分差异(第一列是标签, 因此从 j+1 开始)
	anova_item <- aov(response_b[, j+1] ~ mode, data = response_b)
	# 将结果储存于矩阵相应位置
	TMEbetween[j, 1] <- summary(anova_item)[[1]]$`Pr(>F)`[1]
	}

续表

检验方法	代码示例

MCFA	目的：检验 PBT 与 CBT 下结果的测量不变性
	# 加载所需程序包 -------
	library(lavaan)
	# 模型检验 ----------------
	# (本示例限定所有题目都属于同一个潜在特质)
	# 1. 检验形态等价(即结构不变性)
	# 2. 检验载荷等价(即弱不变性)
	# 3. 检验截距等价(即强不变性)
	# 4. 依次放松每道题目的载荷限制，并将结果储存于 cfa_item
	model <- 'trait =~ item1 + item2 + … + itemN' # 建立模型
	fit1 <- cfa(model, data = response_b, group = "mode") # 形态等价
	fit2 <- cfa(model, data = response_b, group = "mode", group.equal = "loadings") # 载荷等价
	fit3 <- cfa(model, data = response_b, group = "mode",
	group.equal = c("loadings", "intercepts")) # 截距等价
	cfa_item <- matrix(data = NA, nrow = nitem, ncol = 1) # 创建空矩阵
	for (j in 1:nitem){
	# 依次对每一题放松限制
	fit4 <- cfa(model, data = response_b, group = "mode",
	group.equal = c("loadings", "intercepts"),
	group.partial = paste("item", j, "~1", sep = ""))
	# 将结果储存于矩阵相应位置
	cfa_item[j, 1] <- anova(fit3, fit4)$`Pr(>Chisq)`[2]
	}

DIF (SIBTEST)	目的：分析参照组和目标组的结果差异
	# 加载所需程序包 -------
	library(mirt)
	# DIF 检验 -----------------
	# beta_statistic 用于储存检验统计量的结果，并且：
	# $\beta \in (0,0.05)$ 表示不存在 DIF
	# $\beta \in (0.05,0.1)$ 表示存在中等程度 DIF
	# β 大于 0.1 表示存在较严重 DIF (Puhan et al., 2007)
	# suspect 为可能存在 TME 的题目集合
	# anchor 为不存在 TME 的锚题集合
	#(当不指定锚题时，可令除待检题目外的所有题作为锚题集)
	anchor <- c(1, 2, 3) # 设置锚题为第 1、2 和 3 题
	suspect <- c(1:nitem)[-anchor] # 除去锚题，即得到可能存在 DIF 的题目集合
	beta_statistic <- matrix(data = NA, nrow = length(suspect), ncol = 1) # 创建空矩阵
	for (j in 1:length(suspect)){
	# 对每一题进行 DIF 检验
	dif_item <- SIBTEST(response_b[, -1], response_b$mode,
	match_set = anchor, suspect_set = suspect[j])
	# 将结果储存于矩阵相应位置
	beta_statistic[j, 1] <- dif_item$beta[1]
	}